Unit 1 "Date Warehousing"

1. Explain data warehouse architecture.

Ans:

A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise.

Each data warehouse is different, but all are characterized by standard vital components.

Production applicants such as payroll accounts payable product purchasing and inventory control are designed for online transaction processing (OLTP).

Such applicants gather detailed data from day to day operations.

Data warehouse applicants are designed to support the user ad-hoc data requirements, an activity recently dubbed online analytical Processing (OLAP).

Scanned by TapScanner

These include applications such as forecasting, Profiling, Summary reporting, and trend analysis.

Production databases are updated Continuously by either by hand or via OLAP applications.

In contrast, a warehouse database is updated from operational systems periodically, usually during off-hours.

As OLTP data accumulates in production databases, it is regularly extracted, filtered, and then loaded into a dedicated warehouse Server that is accessible to users.

As the warehouse is populated, it must be restructured tables de-normalized, data cleansed of errors and redundancies and new fields and keys added to reflect the needs to the user for Sorting, Combining, and Summarizing data.

Scanned by TapScanner

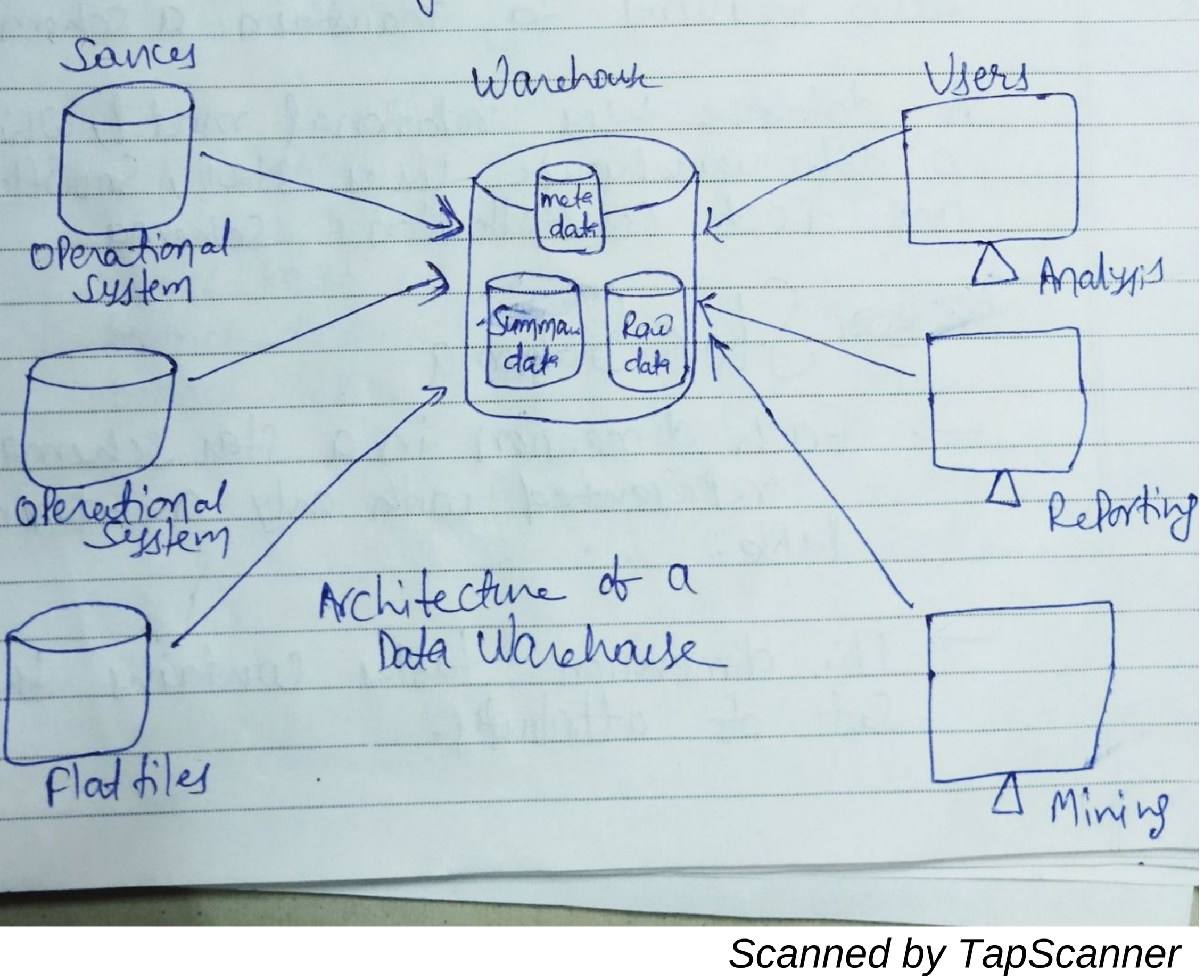
Data warehouses and their architectures vary depending upon the elements of an organization's situation.

Three common architectures are:

Data warehouse Architecture: Basic

Data warehouse Architecture: With Staging Area.

Data warehouse Architecture: With Staging Area and Data Marts



Architecture of a Data Warehouse

Scanned by TapScanner

Roll No. ..... Scholar No. .....

2

What do you mean by Datawarehouse Schema.

Ans.

Schema is a logical description of the entire database.

It includes the name and description of records of all record types including all associated data-items and aggregates.

Much like a database, a data warehouse also requires to maintain a schema.

A database uses relational model, while a data warehouse uses Star, Snowflake and Fact constellation Schema.

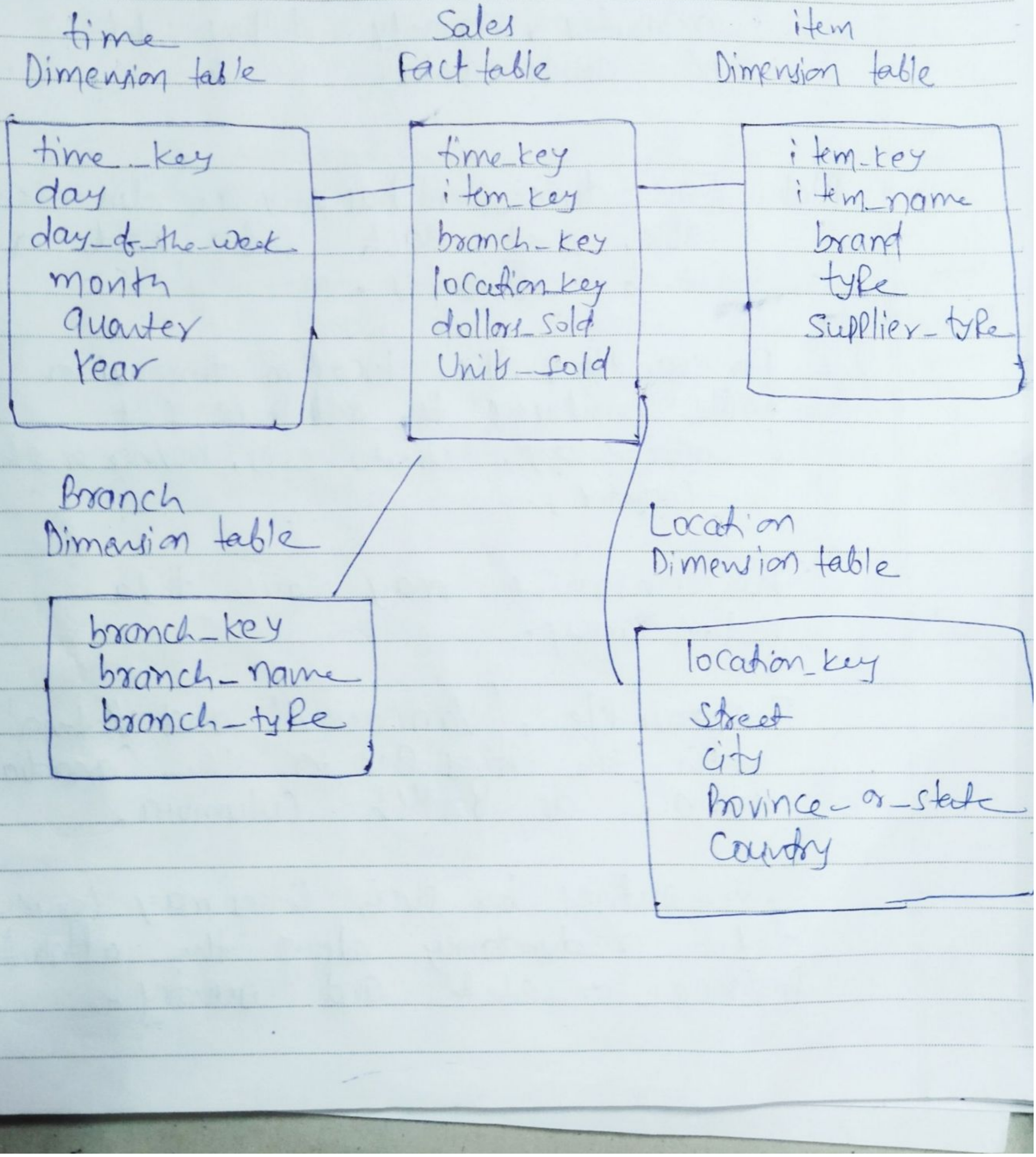
Star Schema

→ Each dimension in a star schema is represented with only one-dimension table.

→ This dimension tables contains the set of attributes.

Scanned by TapScanner

→ The following diagram shows the sales data of a company with respect to the four dimensions, namely time, item, branch, and location.



Scanned by TapScanner

Roll No. ..... Scholar No. .....

→ There is a fact table at the center. It contains the keys to each of four dimensions.

→ The fact table also contains the attributes, namely dollars sold and units sold.

Note- Each dimension has only one dimension table and each table holds a set of attributes.

For example, the location dimension table contains the attribute set { location_key, street, city, Province_or_state, country }.

This constraint may cause data redundancy.

For example, "Vancouver" and "Victoria" both the cities are in the Canadian Province of British Columbia.

The entries for such cities may cause data redundancy along the attributes Province_or_state and country.

Scanned by TapScanner

Snowflake Schema

→ Each dimension in a ~~star schema is~~ represented with only one - dimension table.

→ Some dimension tables in the Snowflake Schema are normalized.

→ The normalization splits up the data into additional tables.

→ Unlike Star-schema, the dimensions table in a snowflake schema are normalized For example, the item dimension table in Star schema is normalized and split into two dimension tables, namely item and Supplier table.

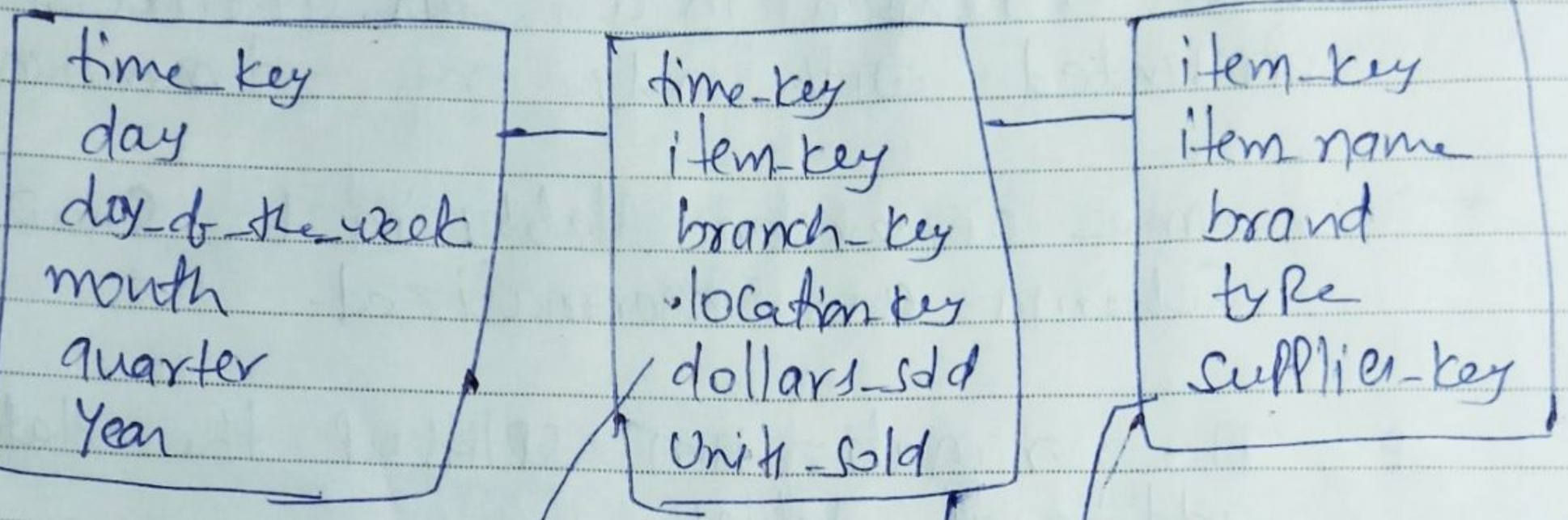
Scanned by TapScanner

Roll No. ..... Scholar No. .....

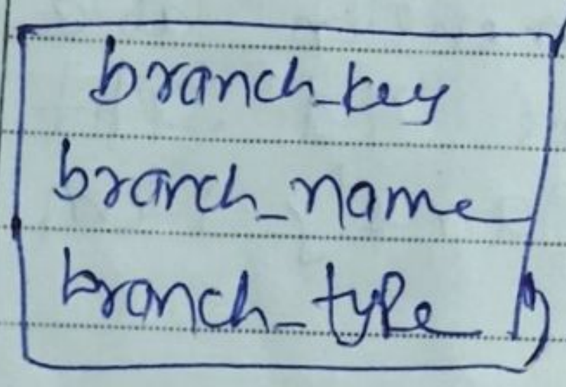
time dimension table

Sales fact table

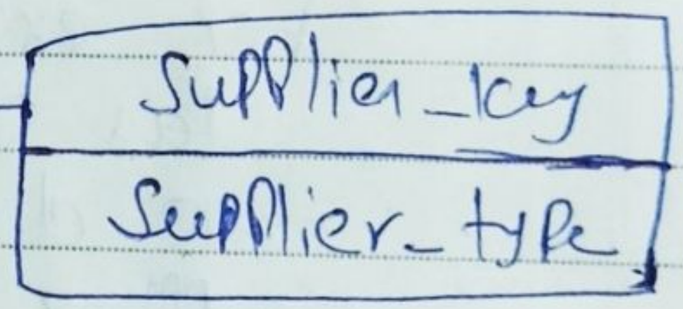
item dimension table



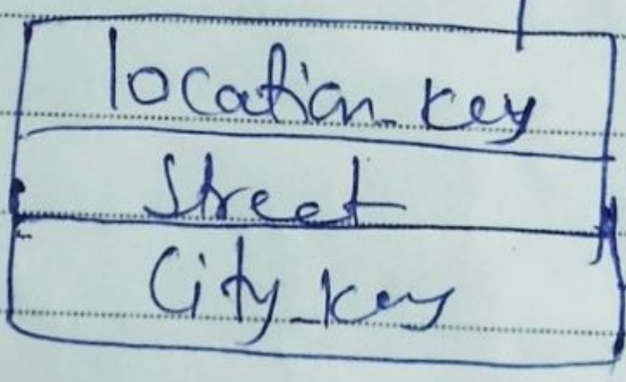
Branch Dimension table



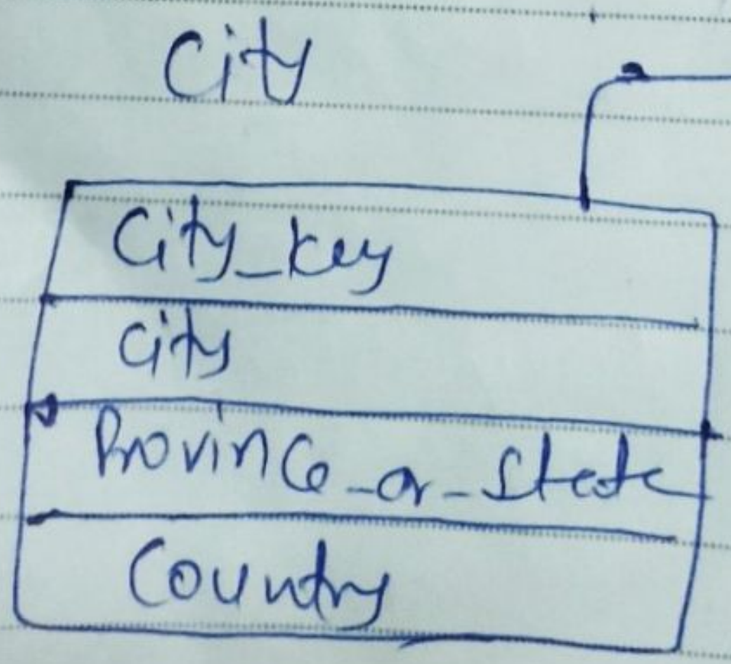
Supplier dimension table



Location Dim. table



City



Scanned by TapScanner

→ Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier key.

→ The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

* Due to normalization in the snowflake schema, the redundancy is reduced and therefore, it becomes easy to maintain and the save storage space.

Fact Constellation Schema

→ A fact constellation has multiple fact tables. It is also known as galaxy schema.

→ The following diagram shows two fact tables, namely Sales and Shipping.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

time
dim_table

Sales
fact table

item
dim.table

Shilling
fact table

time_key
day
day_of_the_week
month
quarter
year

time_key
item_key
branch_key
location_key
dollars_sold
units_sold

item_key
item_name
brand
type
supplier_key

item_key
time_key
Shilller key
branch_key
to_location
dollar_cost
units_sold

branch
dim.table

branch_key
branch_name
branch_type

location
dim.table

location_key
Street
city
Province_or_state
country

Shipper
dim. table

Shipperkey
Shipper_name
locationkey
shipper_type

Scanned by TapScanner

→ The Sales fact table is same as that in the Star Schema.

→ The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.

→ The shipping fact table also contains two measures, namely dollars sold and units sold.

→ It is also possible to share dimension tables between fact tables.

For example, time, item, and location dimension tables are shared between the sales and shipping fact tables.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

3 Explain Multidimensional Data Model.

Ans:

A multidimensional model views data in the form of a data-cube. A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

The dimensions are the perspectives or entities concerning which an organization keeps records.

For example, a shop may create a sales datawarehouse to keep records of the store's sales for the dimension time, item, and location.

These dimensions allow the save to keep track of things, for example, monthly sales of items and the locations at which the items were sold.

Each dimension has a table related to it, called a dimensional table, which describes the dimension further. for example; a dimensional table for an item may contain the attributes item name, brand and type.

Scanned by TapScanner

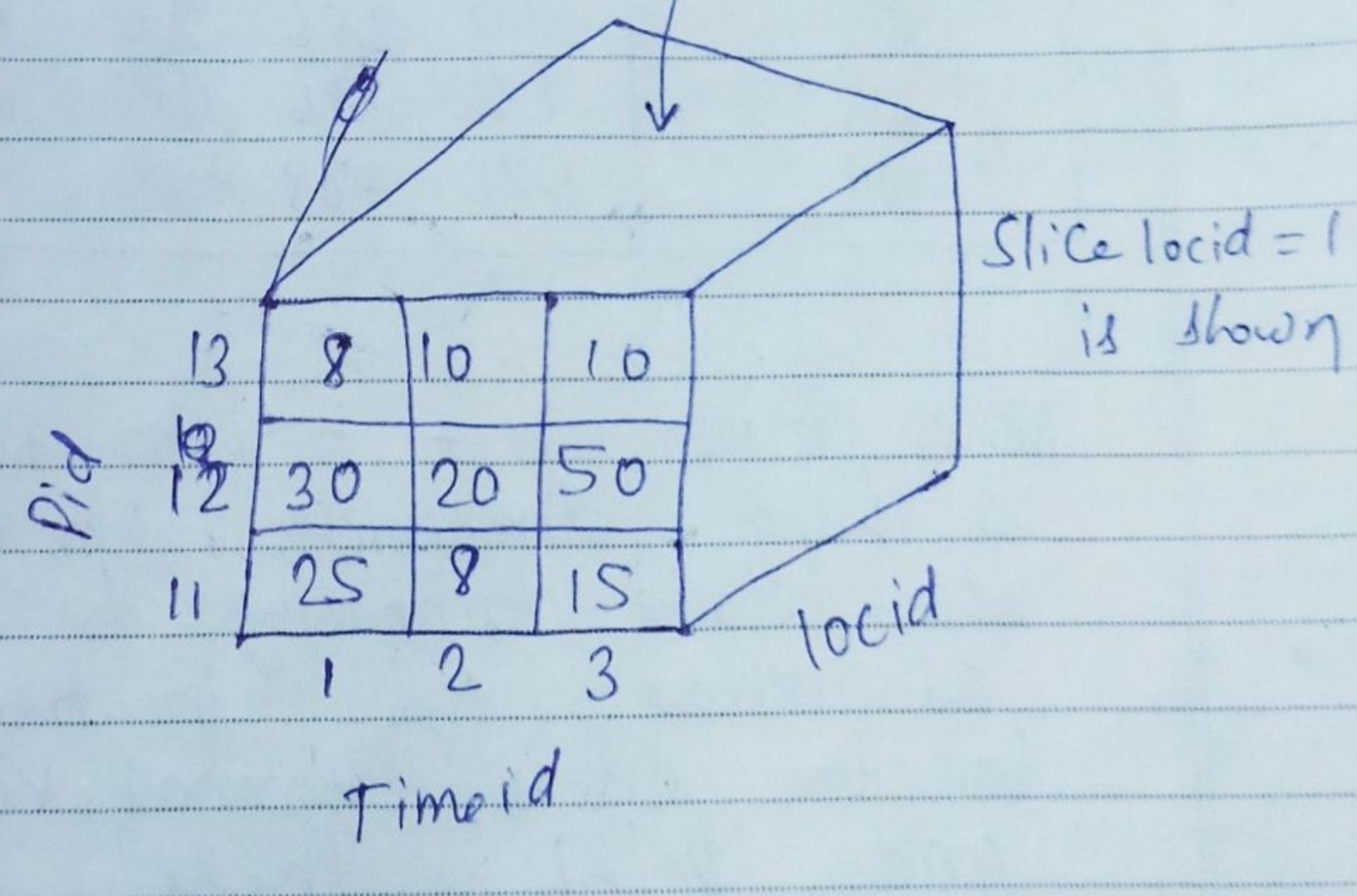
A multidimensional data model is organized around a central theme, for example, sales.

This theme is represented by a fact table.
Facts are numerical measures.
The fact table contains the names of the facts or measures of the related dimensional tables.

Pid	Timeid	locid	Sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

Tabular representation

Multidimensional representation



Consider the data of a shop for items sold Per quarter in city of delhi.
The data is shown in the table.

Scanned by TapScanner

Roll No ..... Scholar No .....

In this 2D representation, the sales of Delhi are shown for the time dimension (organized in quarters) and the item dimension (classified according to the types of an sold item). The fact or measure displayed in rupee_sold (in thousands).

Location = "Delhi"

Time_Quarter	Item(type)			
	Egg	Milk	Bread	Biscuit
Q1	260	508	15	60
Q2	390	256	20	90
Q3	436	396	50	40
Q4	525	483	35	50

Now, if we want to view the sales data with a third dimension, For example, suppose the data according to time and item, as wellas the location is considered for the cities Chennai, Kolkata, Mumbai and Delhi. These 3D data are shown in the delhi.

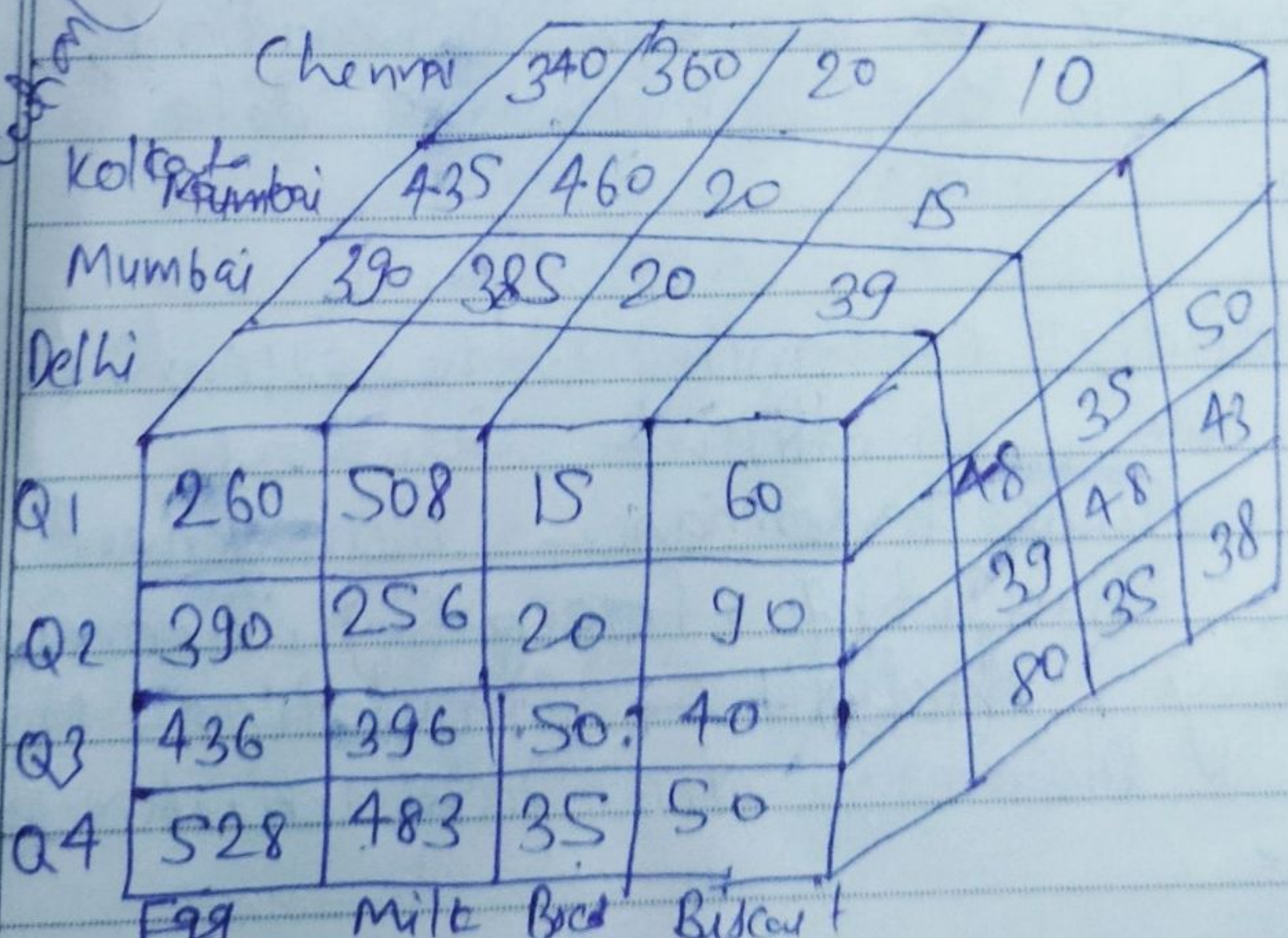
The 3D data of the table are represented as a series of 2D tables

Scanned by TapScanner

	Location "Chennai" item				Location "Kolkata" item				Location "Mumbai" item				Location "Delhi" item			
Time	E	M	B	Bs	E	M	B	Bs	E	M	B	Bs	E	M	B	Bs
Q1	340	360	20	10	435	460	20	15	390	385	20	39	260	508	15	60
Q2	490	490	16	50	389	385	45	35	463	366	25	48	390	256	20	90
Q3	680	583	46	43	684	490	39	48	568	594	36	39	436	396	50	40
Q4	535	694	39	38	335	365	83	35	338	484	48	80	528	483	35	50

Egg = E
Milk = M Biscuit = Bs
Bread = B

Conceptually, it may also be represented by the same data in the form of a 3D data cube, as shown in fig:



Scanned by TapScanner

Roll No. Scholar No.

Unit 2
OLAP Systems

(1) Explain in detail OLAP systems.

OLAP stands for On-Line Analytical Processing. OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast, consistent, interactive access in a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprises as understood by the clients.

OLAP implement the multidimensional analysis of busines information and support the capability for complex estimations, trend analysis and sophisticated data modeling.

It is rapidly enchancing the essential foundation for Intelligent Solutions containing Busines Performance Management, Planning, Budgeting, forecasting, Financial Documenting, Analysis, Simulation-Models, Knowledge Discovery and Data Warehouses Reporting.

Scanned by TapScanner

OLAP enables end-clients to perform ad hoc analysis of record in multiple dimensions, providing the insight and understanding they require for better decision making.

OLAP application are used by a variety of the functions of an organization

Finance and Accounting

- Budgeting
  Activity-based costing
  Financial Performance Analysis
  Financial Modeling

Sales and Marketing

  Sales analysis and forecasting
  Market research analysis
  Promotion Analysis
  Customer Analysis
  Market and customer segmentation

Production

  Production Planning
  Defect analysis

Scanned by TapScanner

Roll No. ..... Scholar No. .....

OLAP cubes have two main purposes. The first is to provide business users with a data model more intuitive to them than a tabular model. This model is called a Dimensional model.

The second purpose is to enable fast query response that is usually difficult to achieve using tabular models

Working of OLAP

OLAP has a very simple concept. It pre-calculates most of the queries that are typically very hard to execute over tabular databases, namely aggregation, joining and grouping.

These queries are calculated during a process that is usually called 'building' or processing of the OLAP cube.

This process happens overnight, and by the time end users get to work - data will have been updated.

Scanned by TapScanner

Dr E.F. Codd, the father of the relational model.

- Multidimensional Conceptual View
- Transparency
- Accessibility
- Consistency Reporting Performance
- Client/Server Architecture
- General Dimensionality
- Dynamic Sparse Matrix Handling
- Multiuser Support
- Unrestricted Cross-dimensional Operations
- Intuitive Data Manipulation
- Flexible Reporting
- Unlimited Dimensions and Aggregation Levels

Scanned by TapScanner

Roll No. ..... Scholar No. .....

2.

Explain
What are OLAP ~~systems~~ Servers ? Explain types of OLAP Servers.

Ans

Online Analytical Processing Server (OLAP) is based on the ~~fundamental~~ multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent and interactive access to information.

Tyes of OLAP Servers.

There are four tyes of OLAP Servers

Relational OLAP (ROLAP)
Multidimensional OLAP (MOLAP)
Hybrid OLAP (HOLAP)
Specialized SQL Server

Scanned by TapScanner

i) Relational OLAP

ROLAP Servers are placed between relation back-end Server and client front end tools.

To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following —

→ Implementation of aggregation navigation logic.
→ Optimization of each DBMS back end.
→ Additional tools and Services.

ii) Multidimensional OLAP

MOLAP uses array-based multidimensional storage engines for multidimensional views of data.

With multidimensional data stores, the ~~Storge~~ Storage utilization may be low if the data set is sparse.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

Therefore, many MOLAP Server use two levels of data storage representation to handle dense and sparse data sets.

iii) Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP.

It offers higher scalability of ROLAP and faster computation of MOLAP.

HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store

iv) Specialized SQL Servers

Specialized SQL Servers provide advanced query language and query processing support for SQL queries over Star and Snowflake Schemas in a read-only environment.

Scanned by TapScanner

3

What do you mean by Backup and Recovery ?

Ans.

Backup refers to storing a copy of original data which can be used in case of data loss.

Backup is considered one of the approaches to data Protection.
Important data of the organization needs to be kept in backup efficiently for Protecting valuable data.

Backup can be achieved by storing a copy of the original data seprately or in a database on storage devices.

There are various types of backups are available like

full backup.
Incremental backup
Local Backup
Mirror backup
etc.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

An example of a Backup can be SnapManager which makes a backup of everything in the database.

Recovery

Recovery refers to restoring lost data by following some Processes.

Even if the data was backed up still lost so it can be recovered by using/ implementation some recovery techniques.

When a database fails due to any reason then there is a chance of data loss, so in that case recovery process helps in improve the reliability of the database.

An example of Recover can be SnapManager which recovers the data from the last transaction.

Scanned by TapScanner

Unit-3
Introduction to Data / Data Mining

1. Explain Data / Data Mining

Ans:

Data

In Computing, data is information that has been translated into a form that is efficient for movement or processing.

Relative to today's computers and transmission media, data is information converted into binary digital form. It is acceptable for data to be used as a singular subject or a plural object.

Data can come in the form of text, observations, figures, images, numbers, graphs or symbols.

For example data might include individual prices, weights, addresses, ages, names, temperatures, dates or distances.

Data is raw form of knowledge and, on its own, doesn't carry any significance or purpose.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

Data Mining :-

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data.

Data mining is also called a knowledge Discovery in Database (KDD).

The knowledge discovery process includes Data cleaning, Data integration, Data Selection, Data Transformation, Data mining, Pattern evaluation and knowledge Presentation.

The Process of extracting information to identify patterns, trends and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data mining.

Scanned by TapScanner

In other words, we can say that Data mining is the Process of investigating hidden Patterns of information to various Perspectives for categorization into useful data, which is collected and assembled In Particular areas such as data warehouse, efficient analysis, data Mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.

Data Mining is the act of automatically Searching for large stores of information to find trends and Patterns that go beyond simple analysis procedures.

Data Mining utilizes complex mathematical algorithms for data segments and evaluates the Probability of future events.

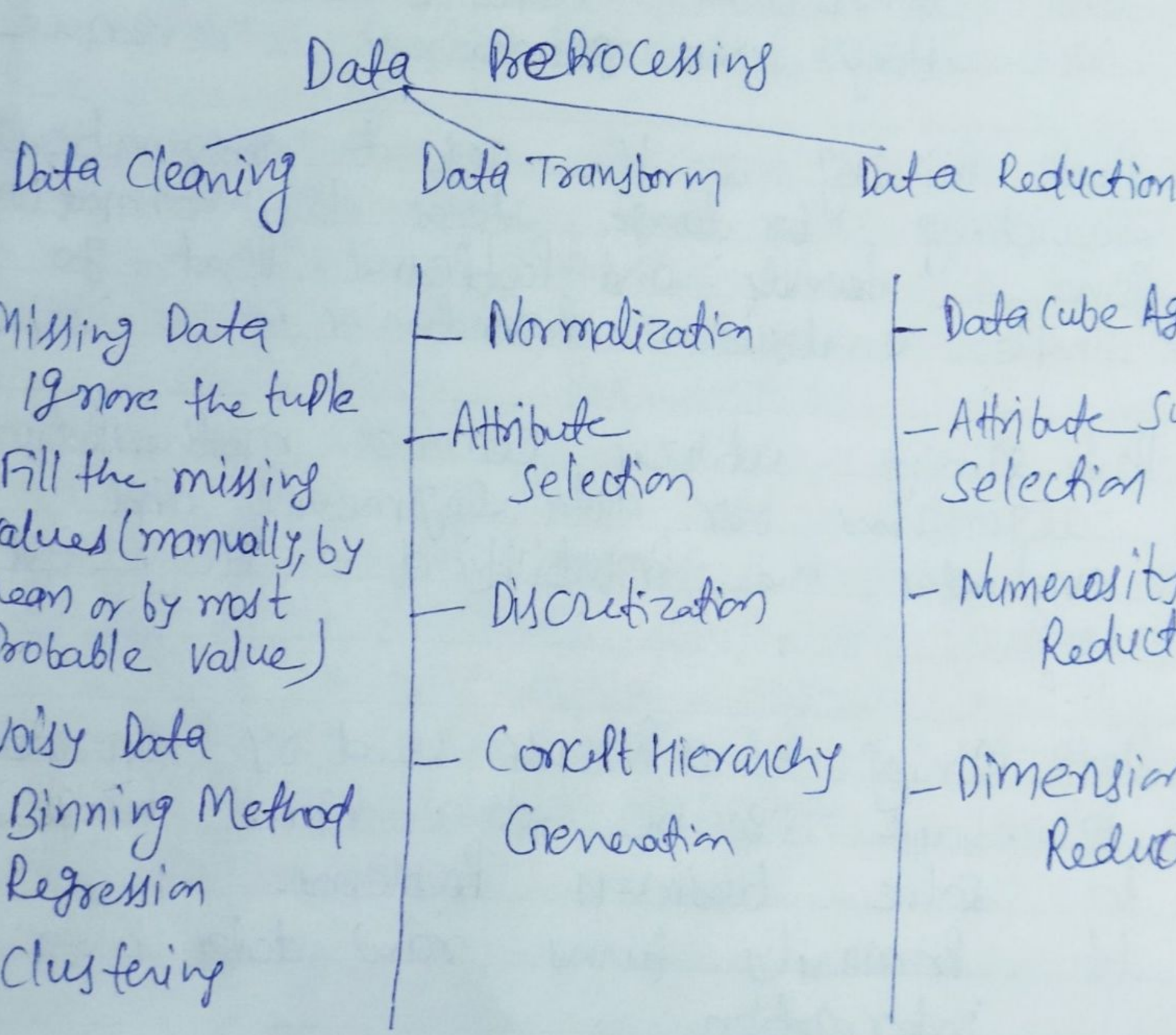
Data Mining is a Process used by organizations to extract specific data from huge databases to solve business Problems.
It Primarily turns raw data into useful information.

Scanned by TapScanner

2. What do You mean by Data Preprocessing.

Ans

Data Preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Scanned by TapScanner

Steps Involved in Data Preprocessing:

1. Data Cleaning

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

a) Missing Data

This ~~missing~~ situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

i) Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

ii) Fill the missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

b) Noisy Data

Noisy data is a meaningless data that can't be interpreted by machines.
It can be generated due to faulty data collection, data entry errors etc.

It can be handled in following ways.

i) Binning Method

ii) Regression

iii) Clustering

2. Data Transformation

This step is taken in order to transform the data in appropriate forms suitable for mining process.

This involves following ways.

a) Normalization

It is done in order to scale the data values in a specified range
(-1.0 to 1.0 or 0.0 to 1.0)

Scanned by TapScanner

b) Attribute Selection

In this strategy, new attributes are constructed from the given set of attributes to hell the mining Procw

c) Discretization:

This is done to replace the raw values of numeric attribute by interval levels or Conceptual levels.

d) Concept Hierarchy Generation.

Here attributes are converted from lower level to higher level in hierarchy. For example -
The attribute "City" can be converted to "Country"

3. Data Reduction

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

In order to get rid of this, we uses data reduction technique.

It aims to increase the storage efficiency and reduce data storage and analysis costs.

3. Explain Fuzzy sets and Fuzzy logic.

Fuzzy Set

i) Fuzzy Set is a set having degrees of membership between 0 and 1.

Fuzzy sets are represented with tilde character ($\sim$).

For example, Number of cars following traffic signals at a particular time out of all cars present will have membership value between [0,1].

ii) Partial membership exists when member of one fuzzy set can also be a part of other fuzzy sets in the same universe

Scanned by TapScanner

iii) The degree of membership or truth is not same as probability, fuzzy truth represents membership in vaguely defined sets.

iv) A fuzzy set $\tilde{A}$ in the universe of discourse, U, can be defined as a set of ordered pairs and it is given by

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x)) \mid x \in X\}$$

1. When the universe of discourse, U, is discrete and finite, fuzzy set $\tilde{A}$ is given by

$$\tilde{A} = \sum_{i=1}^{n} \frac{\mu_{\tilde{A}}(x_i)}{x_i} = \frac{\mu_{\tilde{A}}(x_1)}{x_1} + \frac{\mu_{\tilde{A}}(x_2)}{x_2} + \cdots + \frac{\mu_{\tilde{A}}(x_n)}{x_n}$$

$$\tilde{A} = \int \frac{\mu_{\tilde{A}}(x)}{x}$$

Scanned by TapScanner

Roll No. ..... Scholar No. .....

1. Fuzzy sets also satisfy every property of classical sets.

2. Common Operations on fuzzy sets:

Given two fuzzy sets $A\sim$ and $B\sim$

- Union: Fuzzy Set $C\sim$ is union of Fuzzy Set $A\sim$ and $B\sim$

$$\tilde{C} = \tilde{A} \cup \tilde{B},$$

$$\mu_{\tilde{C}}(x) = \max\left(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\right)$$

- Intersection: Fuzzy Set $D\sim$ is intersection of Fuzzy Sets $A\sim$ and $B\sim$.

$$\tilde{D} = \tilde{A} \cap \tilde{B}$$

$$\mu_{\tilde{D}}(x) = \min\left(\mu_{\tilde{A}}(x), \mu_{\tilde{B}}(x)\right)$$

- Complement: Fuzzy Set $E\sim$ is complemad of Fuzzy Set $A\sim$:

Scanned by TapScanner

$$\tilde{E} = C_{\tilde{A}} X$$

$$\mu_{\tilde{E}}(x) = 1 - \mu_{\tilde{A}}(x)$$

1. Some

$$\mu_A + \mu B(x) = \mu A(x) + \mu B(x) - \mu A(x) . \mu B(x)$$

$$\mu A.B(x) = \mu A(x) . \mu B(x)$$

Bound sum $\mu A \oplus B(x) = \min\{1, \mu A(x) + \mu B(x)\}$

Bound differe.

$$\mu A \ominus B(x) = \max\{0, \mu A(x) - \mu B(x)\}$$

Fuzzy logic

Set: A set is defined as a collection of objects which share certain characteristics.

Classical set

- Classical set is a collection of distinct objects. For example, a set of student passing grades.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

→ Each individual entity in a set called a member or an element of the set.

→ The classical set is defined in such a way that the universe of discourse is splitted into two groups members and non-members.

Hence, In case classical sets, no partial membership exists.

— Let A is a given set. The membership function can be use to define a set A is given by.

$$\mu A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Operations on classical sets:

For two set A and B and Universe X

• Union

$$A \cup B = \{x / x \in A \text{ or } x \in B\}$$

Scanned by TapScanner

- This operation is also called logical OR
- Intersection

$$A \cap B = \{x / x \in A \text{ and } x \in B\}$$

- This operation is also called logical AND
- Complement

$$A' = \{x / x \notin A, x \in X\}$$

- Difference

$$A / B = \{x / x \in A \text{ and } x \notin B\}$$

Properties of classical sets: For two sets A and B and Universe X:

- Commutativity

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

- Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

Scanned by TapScanner

Roll No. ..... Scholar No. .....

- Distributivity :

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

- Idempotency

$$A \cup A = A$$
$$A \cap A = A$$

- Identity

$$A \cup \phi = A$$
$$A \cap X = A$$
$$A \cap \phi = \phi$$
$$A \cup X = X$$

- Transitivity

If $A \subseteq B$ and $B \subseteq C$, then $A \subseteq C$

Scanned by TapScanner

Unit 4
Supervised Learning

1 What do you mean by Supervised Learning?

Ans

Supervised learning is the types of machine learning in which machines are trained using well "labelled" training data, and on basis of that data, machines predict the output

The labelled data means some input data is already tagged with the correct output.

In Supervised learning, the training data provided to the machines work as the Supervisor that teaches the machines to predict the output correctly.

It applies the same concept as a student learns in the supervision of the teacher.

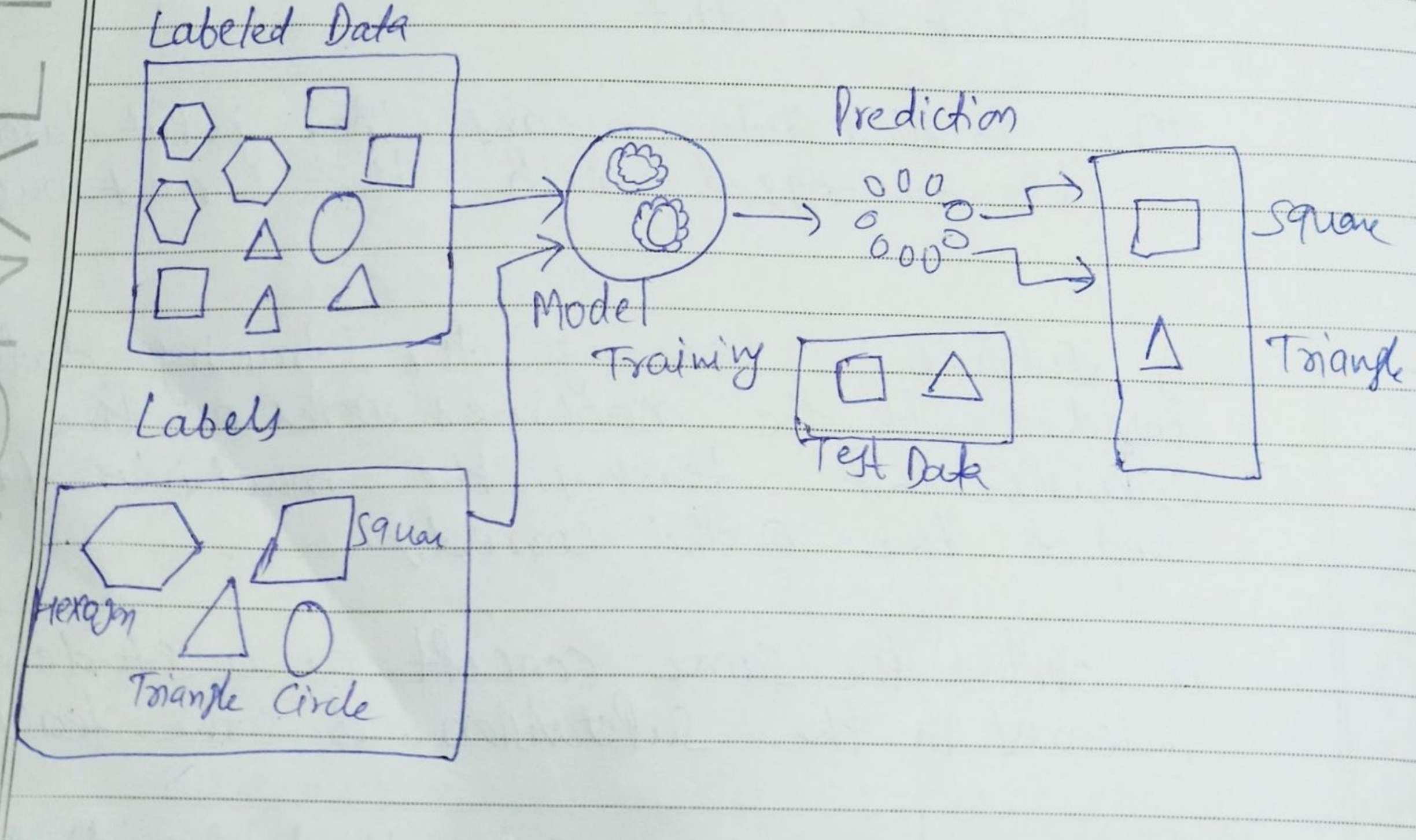
Supervised learning is a process of providing input data as well as correct output data to the machine learning model.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

The aim of a Supervised learning is to find a mapping function to map the input variable (x) with the output variable (y).

In the real world, Supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering etc.



Scanned by TapScanner

2

Explain Statistical-based algorithms.

Ans

A Statistical or data mining is a mathematical expression of certain aspects of the Patterns they find in data.

Different algorithms provide different Perspectives on the complete nature of the Pattern.

There are two types of Statistical-based algorithm which are as follows.

→ Regression
→ Bayesian Classification

i) → Regression

Regression issues deals with the evaluation of an output value located on input values.
When utilized for classification, the input values are values from the database and the output values defines the classes.

Regression can be used to clarify classification issues, but it is used for different

ANSWER SHEETS

applications including forecasting.

The elementary form of regression is simple linear regression that includes only predictor and a prediction.

Regression can be used to implement classification using two various methods which are as follows—

— Division - The data are divided into regions located on class.

— Prediction:- Formulas are created to predict the output class's value

ii) Bayesian Classification

Statistical classifiers are used for the classification. Bayesian classification is based on the Bayes theorem.

Bayesian classifiers view high efficiency and speed when used to high databases.

Scanned by TapScanner

Bayes Theorem

Let X be a data tuple. In the Bayesian method, X is treated as "evidence".

Let H be some Hypothesis, including that the data tuple X belongs to a particularized class C.

The Probability P(H/X) is decided to define the data.
This Probability P(H/X) is the Probability that hypothesis H's influence has given the "evidence" or noticed data tuple X.

P(H/X) is the Posterior Probability of H Conditioned on X. For instance, Consider the nature of data tuples is limited to users defined by the attribute age and income, commonly, and that X is 30 years old users with Rs, 20,000 income.

Assume that H is the hypothesis that the user will Purchase a computer.

Thus P(H/X) reversed the Probability that user X will Purchase a computer

Scanned by TapScanner

Roll No. ..... Scholar No. .....

Given that the user's age and income are acknowledged.

P(H) is the Prior Probability of H.
For instance, this is the Probability that any given user will Purchase a Computer, regardless of age, income ~~are acknowledged~~ or some other data.

The Posterior Probability P(H/X) is located on more data than the Prior Probability P(H), which is free of X.

Likewise, P(X/H) is the Posterior Probability of X conditioned on H.

It is the Probability that a user X is 30 years old and gains Rs. 20,000.

P(H), P(X/H) and P(X) can be measured from the given information.

Bayes theorm suppots a method of computing the Posterior Probability P(H/X), from P(H), P(X/H), and P(X).

Scanned by TapScanner

It is given by

$$P(H/X) = \frac{P(X/H)\ P(H)}{P(X)}$$

2

What do you mean by Decision tree-based algorithms

Ans:-

A decision tree is a non-paramedic Supervised learning algorithm, which utilized for both classification and Regression tasks.

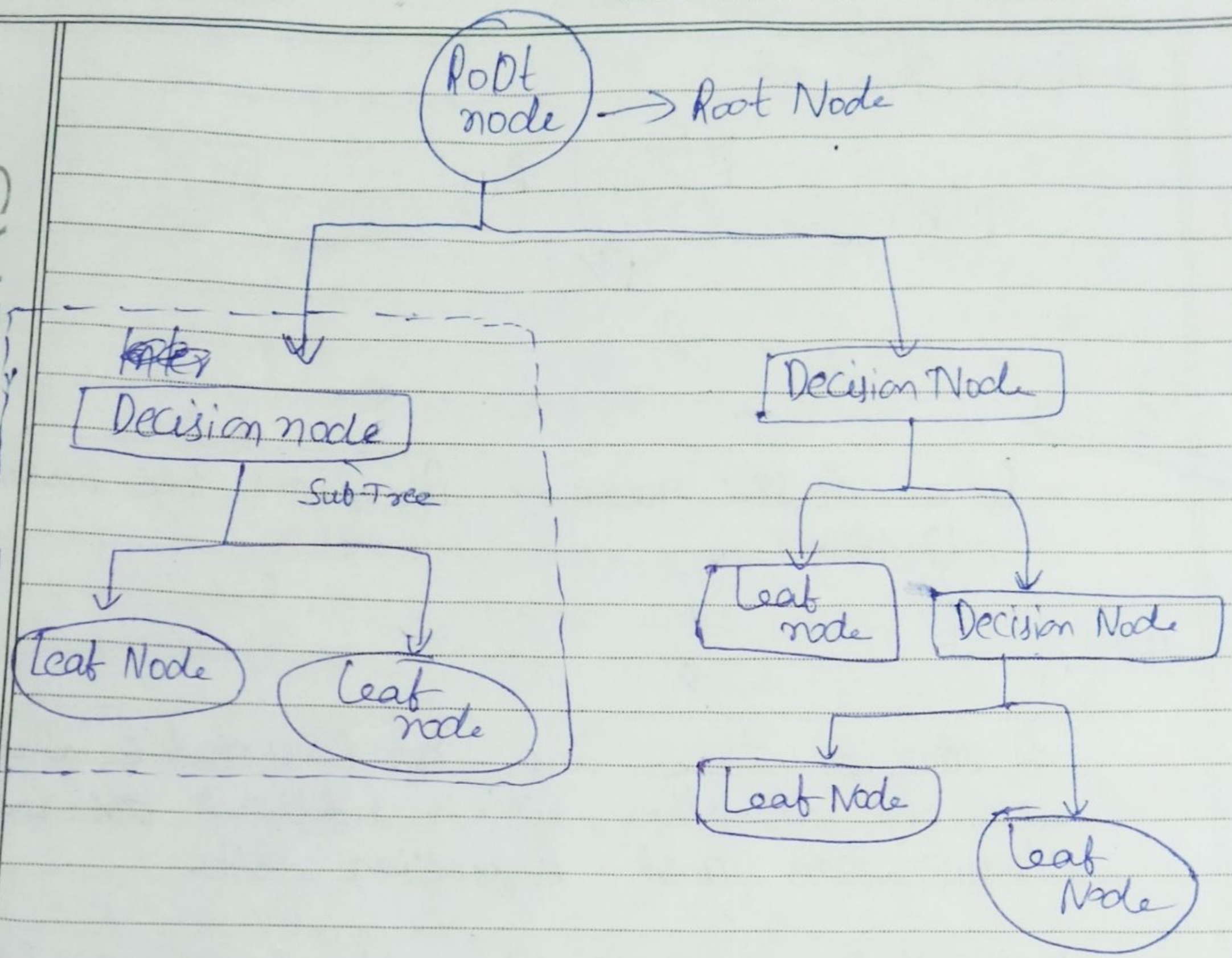
It has a hierarchial, tree structure, which consists of a root node, branches, and internodes and leaf nodes.

In Decision tree, there are two nodes, which are the Decision ~~tree~~ Node and Leaf node.

Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of the those decisions and do not contain any further branches.

Scanned by TapScanner

Roll No. ..... Scholar No. .....



- The decisions or the test are performed on the basis of features of the given dataset.

- It is a graphical representation for getting all the possible solutions to a problem/ decision based on given conditions.

Scanned by TapScanner

- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm

- A decision tree simply asks a question, and based on answer (Yes/No), it further split into subtrees.

- A decision tree can contain categorical Data (YES/NO) as well as numeric data

Reason for using Decision Tree

- Decision tree usually mimic human thinking ability while making a decision, so it is easy to understand.

- Logic behind the decision tree can be easily understood because it shows a tree like structure.

Roll No. ..... Scholar No. .....

RS AL PAPE

Decision Tree Terminologies

- Root node: Root node is from where the decision tree starts.
  It represents the entire dataset, which further gets divided into two or homogeneous sets.

- Leaf Node: Leaf nodes are the final outlet node, and the tree cannot be segregated further after getting a leaf node.

- Splitting:
  Splitting is the process of dividing the decision node/ root node into sub-nodes according to the given conditions.

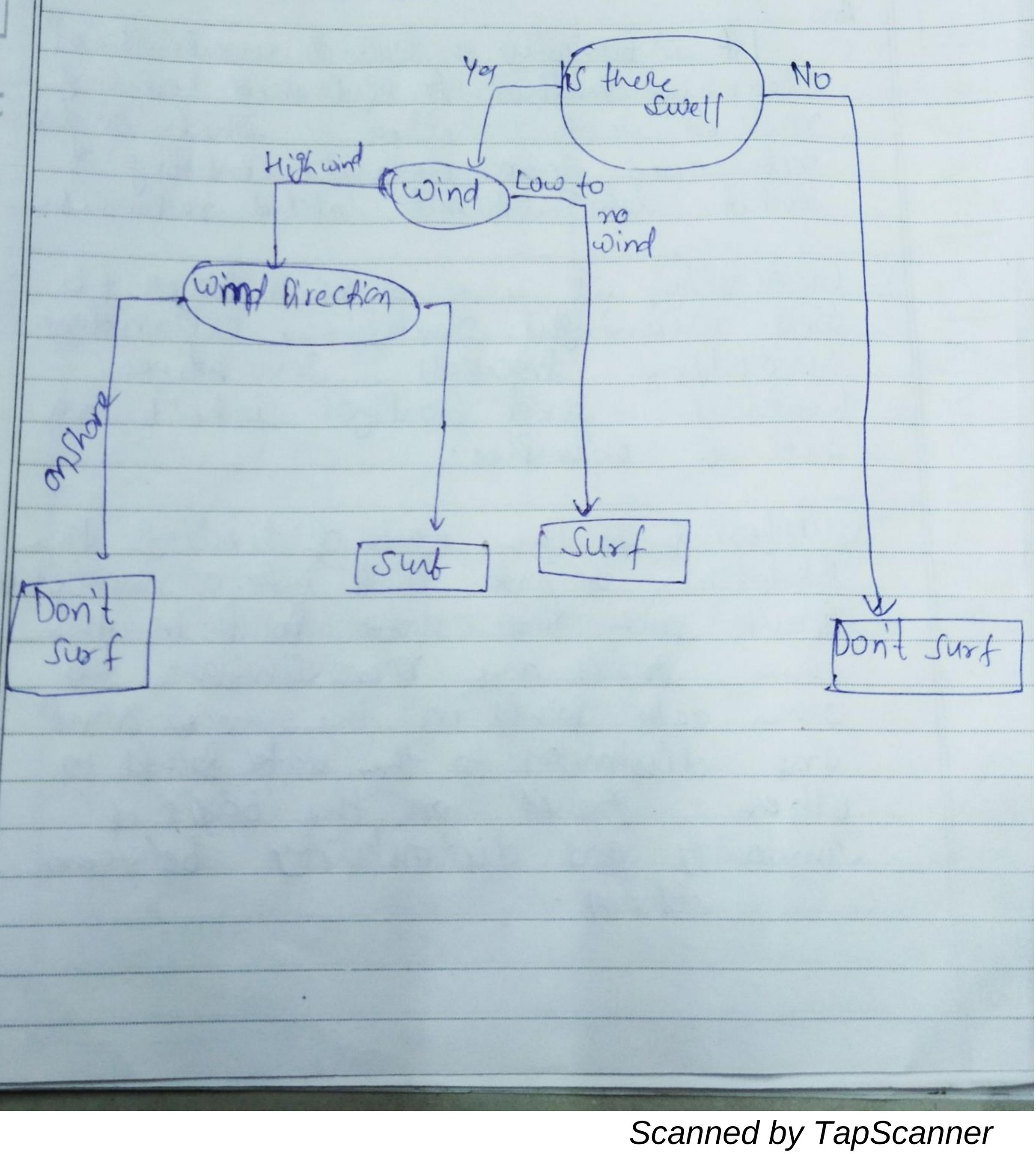
- Branch / Subtree:
  A tree formed by splitting the tree

- Pruning
  Pruning is the process of removing the unwanted branches from the tree.

Scanned by TapScanner

- Parent/child node.

The root node of the tree is called the Parent node, and other nodes are called the child nodes.



Scanned by TapScanner

Roll No. ..... Scholar No. .....

Unit-5
Classification

1. What do you mean by clustering?

Ans

It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses.

Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherit in a set of examples.

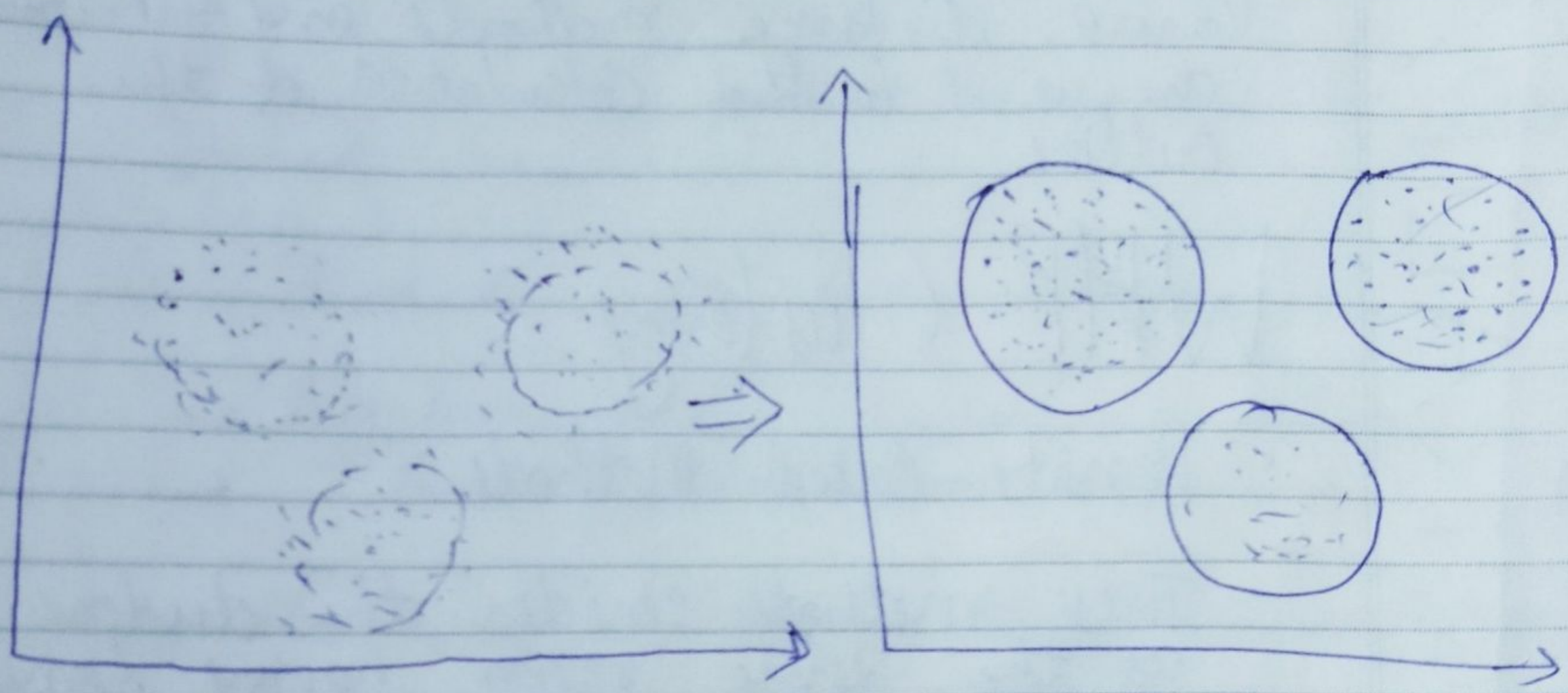
Clustering is the task of dividing the population or data points into a number of group such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups on the basis of similarity and dissimilarity between them.

Scanned by TapScanner

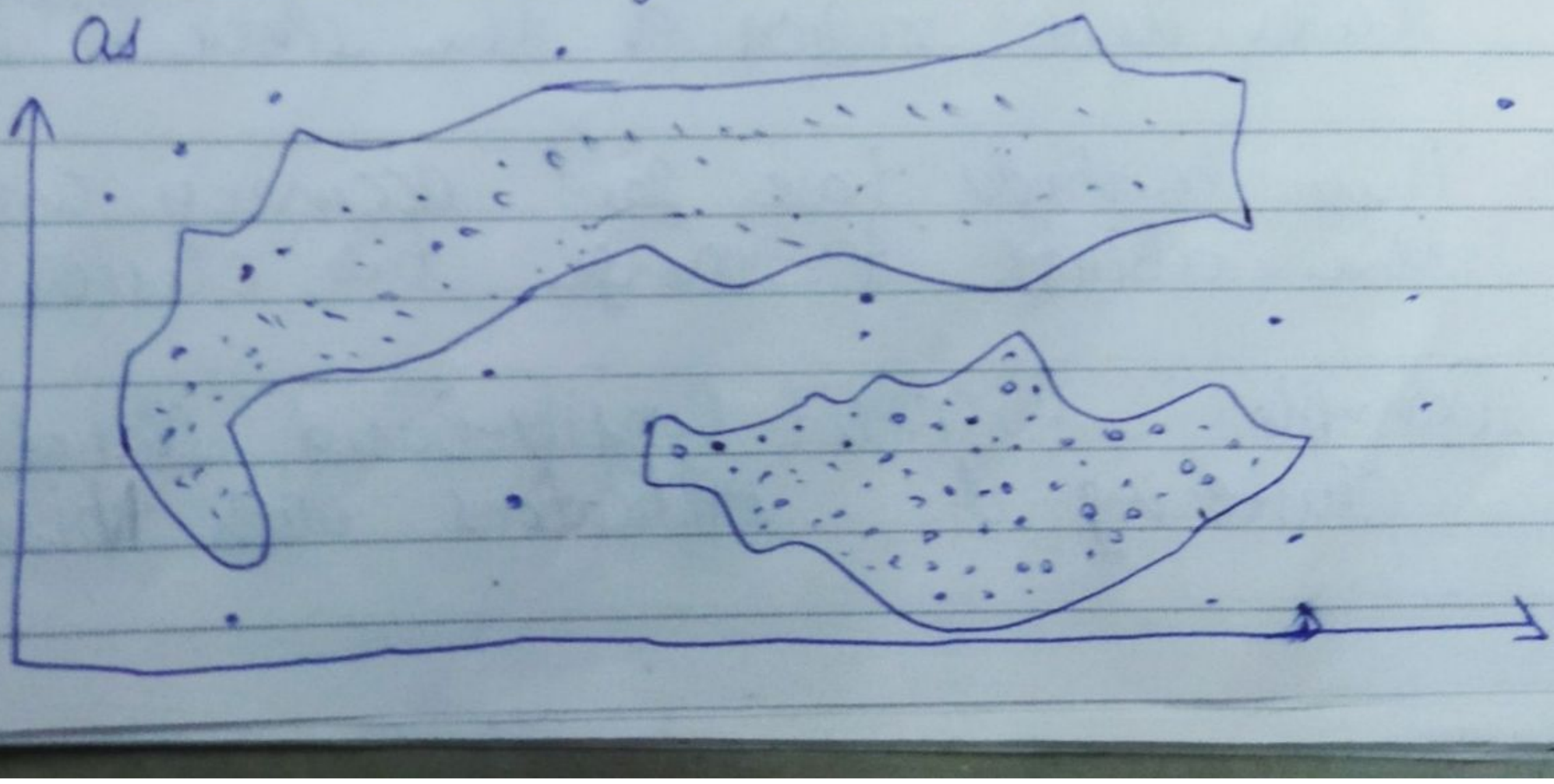
For ex -

The data points in the graph below clustered together can be classified into one single group.

We can distinguish the clusters, and we can identify that there are 3 clusters in the below Pictures



It is not necessary for clusters to be spherical. Such as



Scanned by TapScanner

Roll No. ..... Scholar No. .....

DBSCAN : Density-based Spatial Clustering of Applications with Noise.

These data points are clustered by using the basic concept that the data point lies within the given constraint from the cluster center.

Various distance methods and techniques are used for the calculation of the outliers

WHY Clustering

- Density - Based Methods :

These method consider the clusters as the dense region having some similarities and difference from the lower dense region of the space.

These method have good accuracy and the ability to merge two cluster.

Example: DBSCAN (Density-Based Spatial Clustering of Applications with Noise),

Scanned by TapScanner

OPTICS (Ordering Points to Identify Clustering Structure), etc.

- Hierarchial Based Methods

The clusters formed in this method form a tree type structure based on the hierarchy.

New clusters are formed using the previously formed one.

It is divided into two category

→ Agglomerative (bottom-up approach)
→ Divisive (top-down approach)

example CURE (Clustering Using Representatives),

BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies etc.)

Scanned by TapScanner

Roll No. ..... Scholar No. .....

- Partitioning Methods :

These methods Partition the objects into k clusters and each Partition forms one cluster.

This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example

K-Means, CLARANS (clustering large Applications based upon Randomized Search), etc

- Grid-based Methods :

In this method, the data space is formulated into a finite number of cells that form a grid-like structure.

All the clustering operations done on these grids are fast and independent of the number of data objects example

STING (Statistical Information Grid), wave cluster, CLIQUE (Clustering in Quest), etc

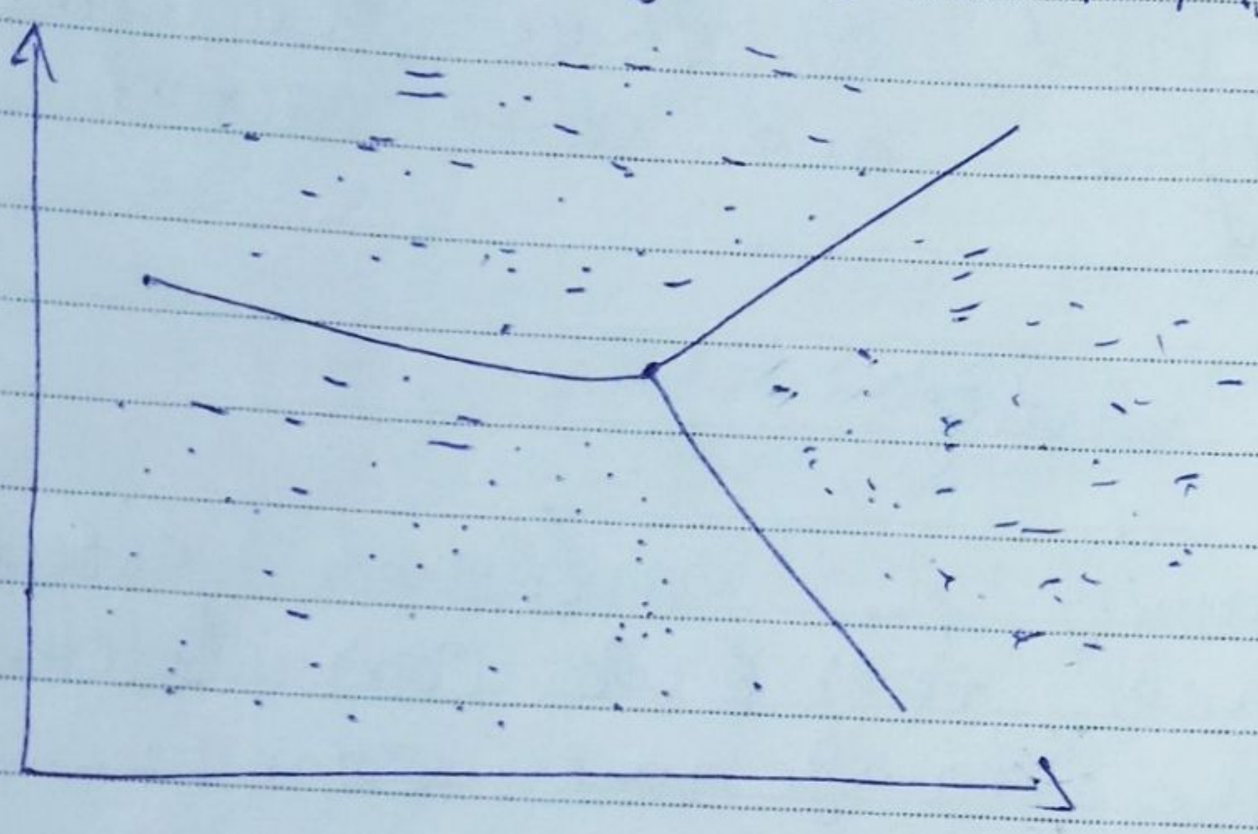
Scanned by TapScanner

Clustering Algorithms

K-Means clustering algorithm –

It is the simplest unsupervised learning algorithm that solves clustering problem.

K-means algorithm Partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



Applications of Clustering in different fields

→ Marketing : It can be used to characterize & discover customer segments for marketing purposes.

→ Biology : It can be used for classification among different species of plants, animals.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

ANSWER PAPERS

- Libraries: It is used in clustering dibbent books on the basis of topics and information.

- Insurance: It is used to acknowledge the customers, their policies and and identifying the brands.

City Planning: It is used to make graphs of houses and to study their values based on their geographical locations and other factors present.

Earthquake Studies:

By learning the earthquake-affected areas can we can determine the dangerous zones.

Scanned by TapScanner

2

Explain Hierarchical algorithms.

Ans: Hierarchical clustering refers to ~~to~~ an unsupervised learning procedure that determines successive clusters based on previously defined clusters.

It works via grouping data into a tree of clusters.

Hierarchial clustering ~~starts~~ starts by treating each data points as an individual cluster.

The endpoint refers to a different set of clusters, where each cluster is different from the other cluster, and the objects within each cluster are same as one another.

There are two types of Hierarchical clustering

- Agglomerative Hierarchical clustering
- Divisive Clustering

Scanned by TapScanner

Roll No. ..... Scholar No. .....

Agglomerative hierarchical clustering

Agglomerative clustering is one of the most common types of hierarchical clustering used to group similar objects in clusters.

Agglomerative clustering is also known as AGNES (Agglomerative Nesting).

In agglomerative clustering, each data point act as an individual cluster and at each step, data objects are grouped in a bottom-up method.

Initially, each data object is in its cluster.

At each iteration, the clusters are combined with different clusters until one cluster is formed.

Scanned by TapScanner

Agglomerative hierarchical clustering algorithm

i) Determine the similarity between individuals and all other clusters.

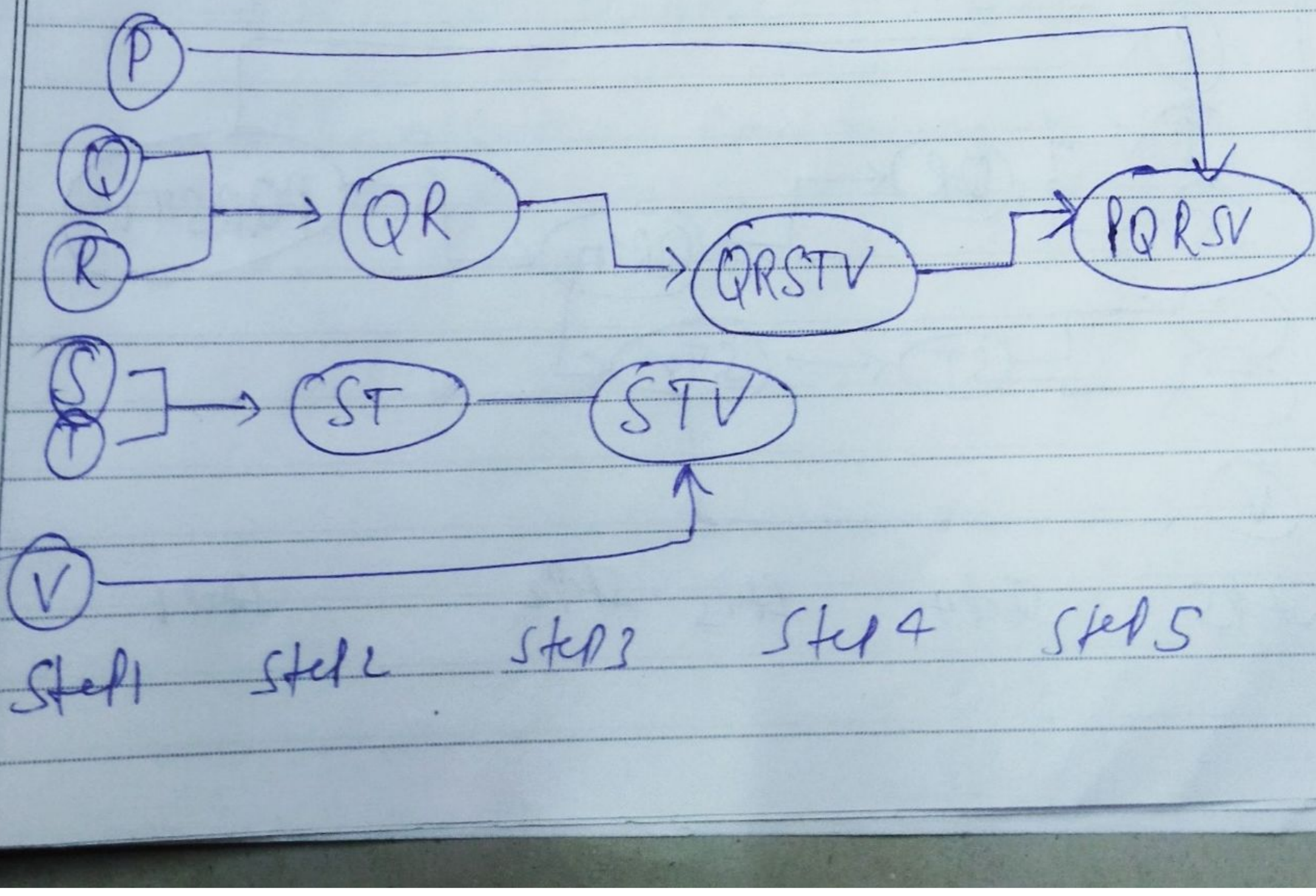
ii) Consider each data point as an individual cluster.

iii) Combine similar clusters.

iv) Recalculate the proximity matrix for each cluster.

v) Repeat step 3 and step 4 until you get a single cluster.

Let's suppose we have six different data point P, Q, R, S, T, V.



Scanned by TapScanner

Roll No. ..... Scholar No. .....

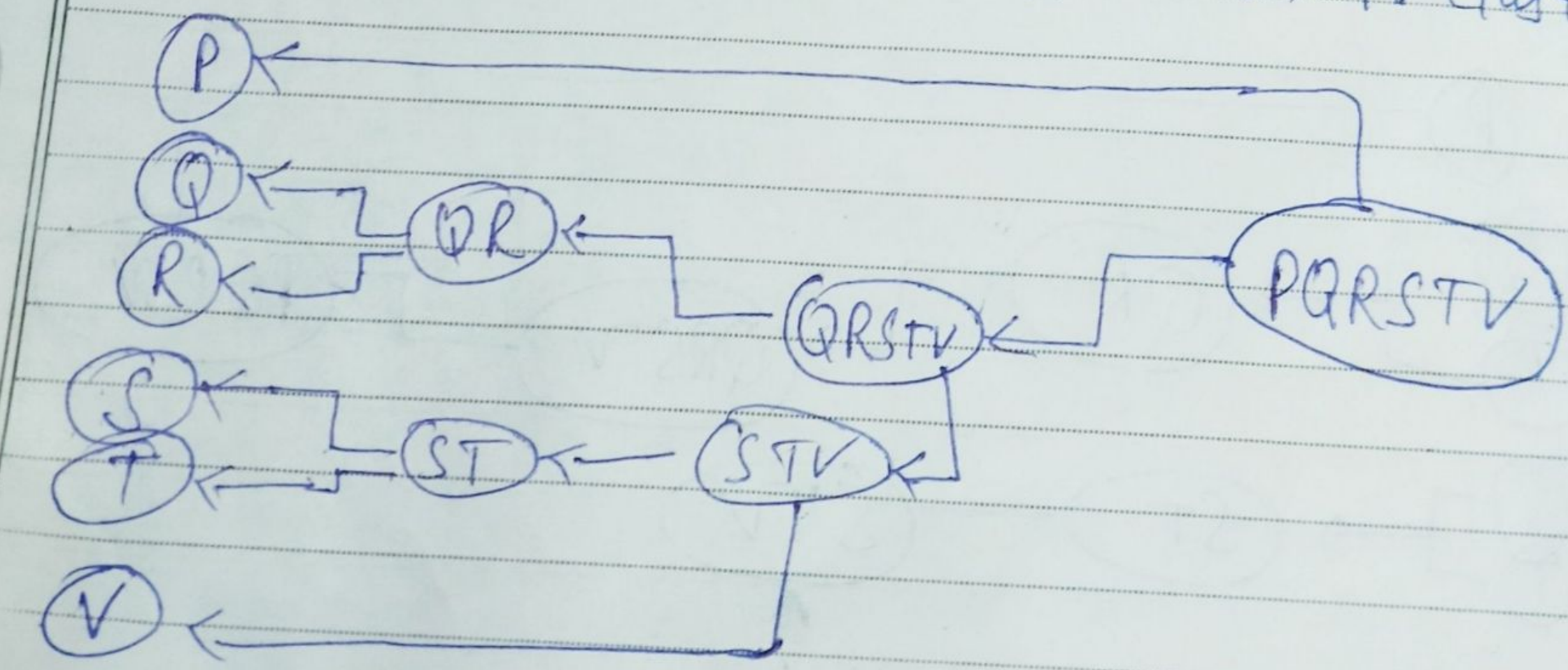
Divisive Hierarchical Clustering

Divisive hierarchical clustering is exactly the opposite of Agglomerative Hierarchical clustering.

In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster.

The separated data points are treated as an individual cluster.

Finally, we are left with N clusters.



Scanned by TapScanner

Advantages

→ It is simple to implement and gives the best output in some cases.

→ It is easy and results in a hierarchy, a structure that contains more information.

→ It does not need us to pre-specify the number of clusters.

Disadvantages

→ It breaks the large clusters.
→ It is difficult to handle different sized clusters and convex shapes.
→ It is sensitive to noise and outliers.

→ The algorithm can never be changed or deleted once it was done previously.

Scanned by TapScanner

Roll No. ..... Scholar No. .....

PEERS VAL PAPER

3 Explain in detail.

a) BIRCH

BIRCH (Balanced Iterative Reducing and Clustering using hierarchies)

is an unsupervised data mining ~~program~~ algorithm that perforns hierarchical clustering over large data sets.

With modifications, it can also be used to accelarate k-means clustering and Gaussian mixture modeling with the expectation - maximization algorithm.

An advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points to produce the best quality clustering for a given set of resources (memory and time constraints).

In most cases, BIRCH only requires a single scan of the database.

Scanned by TapScanner

Basic clustering algorithms like K means and agglomerative clustering are the most commonly used clustering algorithms.

But when performing clustering on very large datasets, BIRCH and DBSCAN are the advanced clustering algorithms used for performing Precise clustering in of large datasets.

Moreover, BIRCH ~~and~~ DBSCAN ~~are~~ are very useful & its easy implementation.

b) DBSCAN

Scanned by TapScanner